



**Tecnológico  
de Monterrey**

Inteligencia artificial avanzada para la ciencia de datos,

Grupo 502

Módulo 5: Estadística Avanzada para Ciencia de Datos

## **REPORTE FINAL**

### **Los peces y el mercurio**

Josue Salvador Cano Martínez | A00829022

Blanca Rosa Ruiz Hernández

04 de diciembre del 2022

## Índice

Resumen	3
Problemática	
Métodos y técnicas estadísticas	
Resultados y conclusiones	
Introducción	3
Problema a resolver	
Importancia del problema	
Análisis de resultados	4
Procedimientos y resultados	
Herramientas estadísticas usadas	
Conclusión	11
Referencias bibliográficas	11
Anexos	12

## **Resumen**

Se pretende determinar, mediante un análisis estadístico, cuáles son los factores (trabajados como variables) que, obtenidos a partir de análisis en lagos de Florida, permitan describir aquello que influye en el nivel de contaminación por mercurio. La primera parte para dar solución a este reto fue un análisis de normalidad de los datos con ayuda de Mardia Test, Anderson Darling Test, Gráfica de contorno, Gráfico QQplot multivariado y distancia de Mahalanobis; la segunda parte consistió en un análisis de componentes principales que ayudaron a definir las variables que impactan en la contaminación, mismas que resultaron ser: la concentración media de mercurio en el tejido muscular del grupo de peces estudiados en cada lago, el máximo de la concentración de mercurio en cada grupo de peces y estimación de la concentración de mercurio en el pez de 3 años.

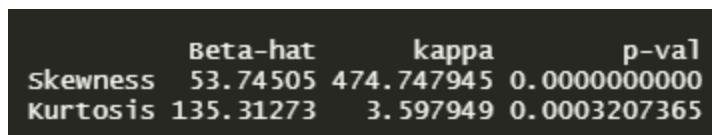
## **Introducción**

La contaminación por mercurio de peces en el agua dulce comestible es una amenaza directa contra nuestra salud. Se llevó a cabo un estudio reciente en 53 lagos de Florida con el fin de examinar los factores que influían en el nivel de contaminación por mercurio. A partir de dicho estudio se determinaron 12 variables que describen las condiciones de cada lago y cuyo fin es determinar las causas de la contaminación para así entender cómo erradicar el problema.

La importancia del problema radica en el impacto directo que tiene sobre el ser humano que se alimenta de los peces de dichos lagos o que incluso su consumo de agua se ve abastecido de los mismos. Resulta necesario determinar las principales causas de la alta concentración de mercurio en dichos recursos naturales con el fin de buscar implementar estrategias que reduzcan sus niveles de concentración y así se garantice la salud de la población y la conservación del medio ambiente.

## Análisis de resultados.

### *Mardia's Test.*



	Beta-hat	kappa	p-val
Skewness	53.74505	474.747945	0.0000000000
Kurtosis	135.31273	3.597949	0.0003207365

**Imagen 1.** Prueba de Mardia. Fuente: Elaboración propia.

Tomando en consideración la hipótesis:

$H_0$ : Las variables siguen una distribución normal multivariable

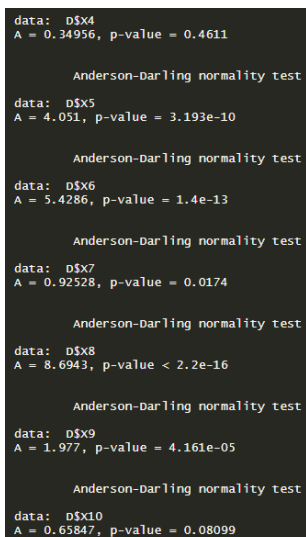
$H_1$ : Las variables no siguen una distribución normal multivariable

Y la regla de decisión con un nivel de significancia del 5%:

Se rechaza  $H_0$  si: el valor  $p$  es menor a  $\alpha = 0.05$

La conclusión es que ya que el  $p$ -value de la curtosis es 0.0003 y del sesgo 0.0000, menores que Alpha; la hipótesis nula se rechaza, por lo no se tiene evidencia para decir que las variables del set de datos siguen una distribución multivariable.

### *Anderson-Darling Test.*



```
data: D5X4
A = 0.34956, p-value = 0.4611

Anderson-Darling normality test
data: D5X5
A = 4.051, p-value = 3.193e-10

Anderson-Darling normality test
data: D5X6
A = 5.4286, p-value = 1.4e-13

Anderson-Darling normality test
data: D5X7
A = 0.92528, p-value = 0.0174

Anderson-Darling normality test
data: D5X8
A = 8.6943, p-value < 2.2e-16

Anderson-Darling normality test
data: D5X9
A = 1.977, p-value = 4.161e-05

Anderson-Darling normality test
data: D5X10
A = 0.65847, p-value = 0.08099
```

**Imagen 2.** Prueba de Anderson-Darling. Fuente: Elaboración propia.

Tomando en consideración la hipótesis:

$H_0$ : Los datos siguen una distribución normal

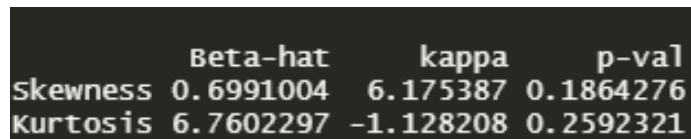
$H_1$ : Los datos no siguen una distribución normal

Y la regla de decisión con un nivel de significancia del 5%:

Se rechaza  $H_0$  si: el valor p es menor a  $\alpha = 0.05$

La conclusión es que ya que el p-value de X4 es 0.46 y de X10 es 0.08, mayores que Alpha; la hipótesis nula no se rechaza, por lo que ambas variables siguen una distribución normal.

*Mardia's test a variables que presentan normalidad.*



	Beta-hat	kappa	p-val
skewness	0.6991004	6.175387	0.1864276
kurtosis	6.7602297	-1.128208	0.2592321

**Imagen 3.** Prueba de Mardia a datos normales. Fuente: Elaboración propia

Tomando en consideración la hipótesis:

$H_0$ : Las variables siguen una distribución normal multivariable

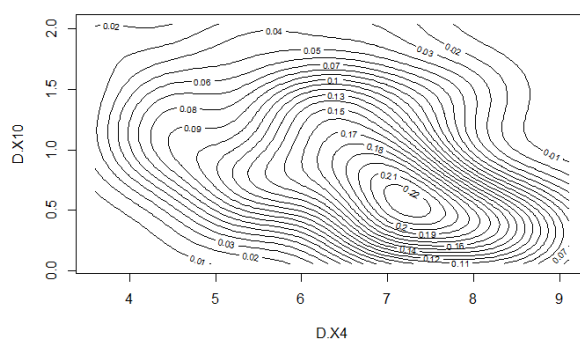
$H_1$ : Las variables no siguen una distribución normal multivariable

Y la regla de decisión con un nivel de significancia del 5%:

Se rechaza  $H_0$  si: el valor p es menor a  $\alpha = 0.05$

La conclusión es que ya que el p-value de la curtosis es 0.25 y del sesgo 0.18, mayores que Alpha; la hipótesis nula no se rechaza, por lo que las variables X4 y X5 siguen una distribución normal multivariable.

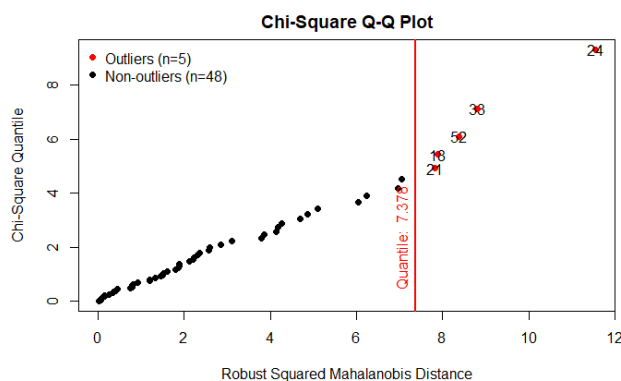
*Gráfica de contorno.*



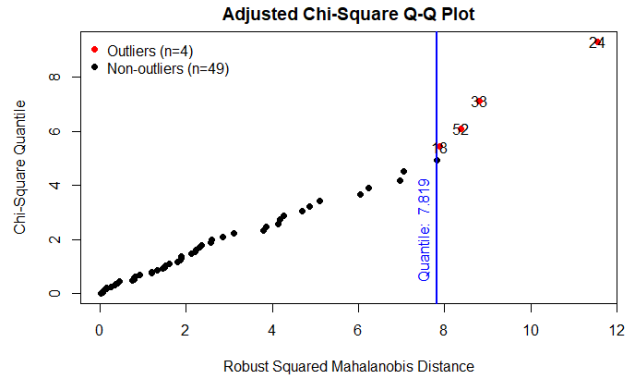
**Gráfica 1.** Gráfica de contorno de la normal multivariada. Fuente: Elaboración propia.

Las regiones con valores más altos se pueden apreciar a partir de los niveles de contorno, mismos que revelan un pico centrado en aproximadamente 7 para el valor de X4 y en 0.5 para el valor de X10, las puntuaciones en esta región pico son superiores a 0.2

*Datos atípicos o influyentes en la normal multivariada*



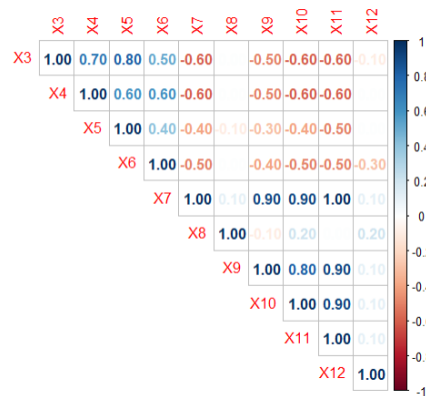
**Gráfica 2.** Gráfica de datos atípicos usando distancia de Mahalanobis. Fuente: Elaboración propia.



**Gráfica 3.** Gráfica de datos atípicos usando distancia de Mahalanobis ajustada. Fuente: Elaboración propia.

De las gráficas obtenidas, la distancia de Mahalanobis declara 5 observaciones como valor atípico multivariado, mientras que la distancia Mahalanobis ajustada declara 4

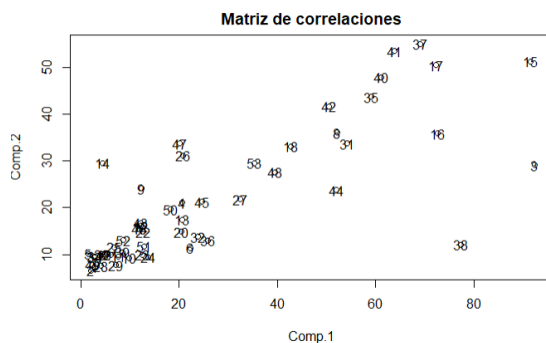
#### *Matriz de correlaciones*



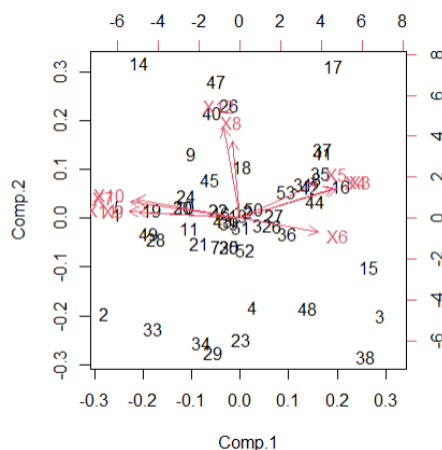
**Imagen 4.** Matriz de correlaciones de datos. Fuente: Elaboración propia

Es adecuado el uso de componentes principales para analizar la base debido a que esto permite identificar aquellas variables que tienen un mayor peso en la contaminación por mercurio de peces en el agua dulce comestible.

## *Análisis de componentes principales*



**Gráfica 4.** Matriz de correlaciones. Fuente: Elaboración propia.

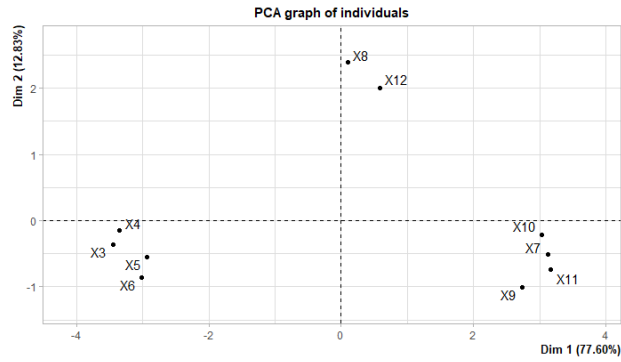


**Gráfica 5.** Componentes principales. Fuente: Elaboración propia.

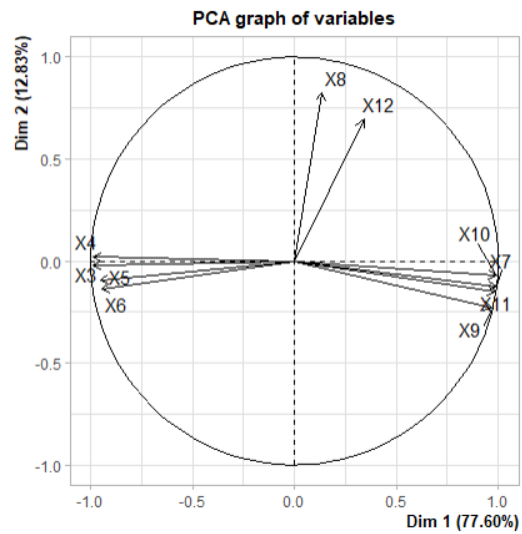
La matriz de correlaciones permite identificar dos componentes que agrupan las variables: la base 1, donde el porcentaje de proporción de varianza explicada es de 77.60% y la base 2 donde el porcentaje es 12.83%. Este número de componentes principales explica poco más del 90% la exactitud, por lo que resulta ideal recurrir al uso de uno de ellos para reducir la dimensión de la base.

*Vectores asociados a las variables y las puntuaciones de las observaciones de las dos primeras componentes*

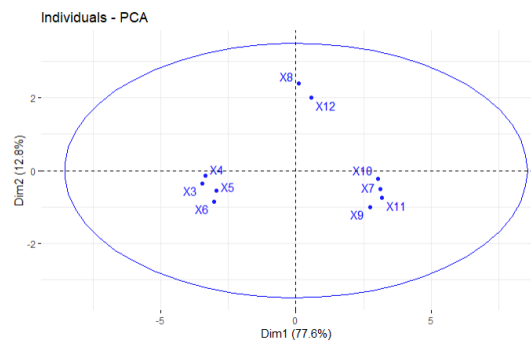




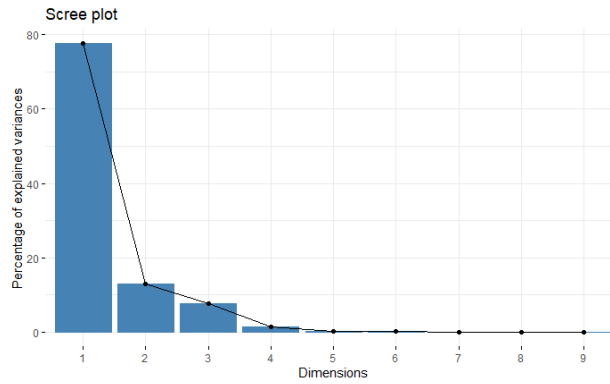
**Gráfica 6.** Componentes principales y sus variables. Fuente: Elaboración propia.



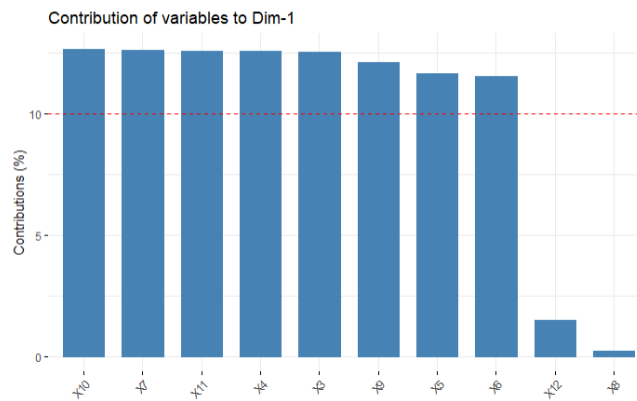
**Gráfica 7.** Componentes principales y sus variables. Fuente: Elaboración propia.



**Gráfica 8.** Componentes principales y sus variables. Fuente: Elaboración propia.



**Gráfica 9.** Porcentajes de varianza explicada por componente. Fuente: Elaboración propia.



**Gráfica 10.** Contribución de las variables al primer componente. Fuente: Elaboración propia.

En el primer gráfico se puede notar el porcentaje de proporción de varianza explicada que tiene cada uno de los dos primeros componentes; se puede notar que para el primero es de 77.60% y para el segundo 12.83%. También, permite definir las variables que tienen una mayor influencia en cada una de ellas. El segundo y tercer gráfico permite entender el mismo resultado descrito en el punto anterior, con una visualización distinta (agrupada por cuadrantes). El penúltimo gráfico permite visualizar la proporción de varianza explicada en cada componente, donde se puede demostrar que el primero es el que mayormente explica los datos. El último gráfico permite visualizar el porcentaje de contribución (peso) de cada variable en el primero componente.

## Conclusión

Los principales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida son: la concentración media de mercurio (parte por millón) en el tejido muscular del grupo de peces estudiados en cada lago, el máximo de la concentración de mercurio en cada grupo de peces y estimación (mediante regresión) de la concentración de mercurio en el pez de 3 años. La normalidad encontrada en un grupo de variables detectadas ayuda dentro de este estudio a que se efectúen con mayor precisión los análisis estadísticos de componentes principales. La distribución normal sirve para conocer la probabilidad de encontrar un valor de la variable que sea igual o inferior a un cierto valor, conociendo la media, la desviación estándar, y la varianza de un conjunto de datos en sustituyéndolos en la función que describe el modelo. Los componentes principales se basan en la proporción de varianza explicada, misma que, como se mencionó antes, es positivamente impactada cuando se cuentan con datos normalizados. Así mismo, los componentes principales permiten determinar aquellas variables que explican de mejor manera, para este caso en específico, la contaminación por mercurio de peces en el agua dulce comestible.

## Referencias

- Amat Rodrigo, J. (2017, 06). *RPubs*. Retrieved from [https://rpubs.com/Joaquin\\_AR/287787](https://rpubs.com/Joaquin_AR/287787)
- Finnstats. (2021, 11 09). *R-Bloggers*. Retrieved from <https://www.r-bloggers.com/2021/11/anderson-darling-test-in-r-quick-normality-check/>
- R Coder. (2022). *R Charts*. Retrieved from <https://r-charts.com/es/correlacion/contour-ggplot2/#:~:text=Se%20puede%20crear%20un%20gr%C3%A1fico,las%20variables%20x%20e%20y%20.&text=Es%20posible%20incrementar%20o%20disminuir%20el%20n%C3%BAmero%20de%20niveles%20con%20bins%20>
- Torres Manzaner, E. (2022). Retrieved from <https://torres.epv.uniovi.es/centon/qqplots.html>
- Zhou, M., & Shao, Y. (2014). *Search r*. Retrieved from <https://search.r-project.org/CRAN/refmans/mvnormalTest/html/mardia.html>

## **Anexos**

Documento de análisis en R:

[A00829022\\_Portafolio\\_TC3007C.502/A00829022\\_MomentoDeRetroalimentacion\\_Modulo5ProcesamientoDeDatosMultivariados\\_PortafolioImplementacion.Rmd at main · JosueCano143/A00829022\\_Portafolio\\_TC3007C.502 \(github.com\)](#)